



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

**Scoring System for Predicting
Functional Outcome at 3-months
Based on Linked Big Data after Acute
Ischemic Stroke: S-SMART score**

연계 빅데이터 기반의 급성 뇌경색 환자의 3 개월
후 기능적 예후 예측 점수 체계: S-SMART

score

2019 년 8 월

서울대학교 대학원

의학과 중개의학 뇌신경과학 전공

김 태 정

Abstract

Scoring System for Predicting Functional Outcome at 3-months Based on Linked Big Data after Acute Ischemic Stroke: S-SMART score

Tae Jung Kim

Translational Medicine, Neuroscience, School of Medicine

The Graduate School

Seoul National University

Background and purpose: Linkage of public healthcare data is useful in stroke research because patients may transition between different sectors of the health system before, during, and after stroke. Prediction of outcome after stroke may help clinicians provide effective management and plan long-term care. We aimed to develop and validate a risk score for predicting functional outcome available for hospitals after ischemic stroke using linked big data.

Methods: Acute stroke patients (n=65,311) with claim data suitable for linkage were included in the Clinical Research Center for Stroke (CRCS) registry during 2006-2014. We linked the CRCS registry with national health claim databases in the Health Insurance Review and Assessment Service (HIRA) using 6 common variables: birth date, gender, provider identification, receiving year, receiving number, and statement serial number in the benefit claim statement. For matched records, linkage accuracy was evaluated using differences between hospital visiting

date in the CRCS registry and the commencement date for health insurance care in HIRA. Among the linked data, a total of 22,005 patients with acute ischemic stroke from the CRCS-HIRA linked data between July 2007 and December 2014 were included in the derivation group. We assessed functional outcomes using a modified Rankin scale (mRS) at 3 months after ischemic stroke. We identified predictors related to good 3-month outcome ($\text{mRS} \leq 2$) and developed a score using logistic regression coefficients. The model was validated in two validation (geographic and temporal) groups. Prediction model performance was assessed by the area under the receiver operating characteristic curve (AUC).

Results: The linkage accuracy was 94.4% in the linked data. Stroke Severity, Sex (gender), stroke Mechanism, Age, pre-stroke mRS, and Thrombolysis/thrombectomy treatment were identified as predictors of the S-SMART score (total 34 points) for predicting functional outcome after stroke using linked big data. The AUC of the prediction score was 0.805 (0.798–0.811) in the derivation group for 3-month functional outcome. The AUCs of the model were 0.812 (0.795–0.830) for the geographic external validation group and 0.812 (0.771–0.854) for the temporal external validation group.

Conclusion: We could establish big data on stroke by linking CRCS registry and HIRA records, using claims data without personal identifiers. Moreover, the S-SMART score is a valid, externally reliable tool to predict functional outcome following ischemic stroke. This prediction model may assist estimation of functional outcome after stroke so as to determine care plans after stroke.

Keywords: big data, data linkage, ischemic stroke, prognosis, prediction score

Student number: 2014-30679

CONTENTS

Abstract	i
List of Tables	iv
List of Figures	v
Introduction	1
Methods	3
Results.....	10
Discussion	28
References.....	33
Abstract in Korean	39

LIST OF TABLES

Table 1. Baseline characteristics of matched cases according to linkage status.....	11
Table 2. Baseline characteristics of derivation and geographic validation groups..	13
Table 3. Baseline characteristics of derivation and temporal validation group.....	15
Table 4. Full version scoring system for prediction of 3-month outcome in ischemic stroke patients	17
Table 5. S-SMART score for prediction of 3-month outcome in ischemic stroke patients	20
Table 6. Prognostic scores for predicting outcome in stroke patients	29

LIST OF FIGURES

Figure 1. Flow diagram of establishing linkage dataset	5
Figure 2. Flow diagram of included cases for developing prediction score system .	7
Figure 3. Receiver operating characteristic (ROC) curves of full version prediction score for functional outcome in derivation and external validation groups.....	19
Figure 4. Receiver operating characteristic (ROC) curves of the S-SMART score for functional outcome in derivation and external validation groups	21
Figure 5. Association of prediction scores with good functional outcome	22
Figure 6. Calibration plots of S-SMART score for predicting functional outcome in two external validation groups	24
Figure 7. Calibration plots of full version score for predicting functional outcome in two external validation groups.	25
Figure 8. Comparing prediction power of S-SMART score to THRIVE score.....	26

Introduction

Stroke is a devastating disease for patients and families and a leading cause of disability. More than 50% of stroke patients continue to experience motor deficits, associated with diminished quality of life.^{1,2} Prediction of outcome after stroke may help clinicians provide effective stroke management, as well as plan discharge and long-term care. Previously, several ischemic stroke outcome prediction scores have been developed that predict outcomes, including mortality, risk of hemorrhage after thrombolysis, and functional outcomes.³⁻¹⁵ However, there is as yet no agreement on any standardized score based on the data available during the acute stage, and those that exist display differences in terms of prognostic accuracy.¹⁶ Moreover, some scores required neurological symptoms or imaging information, and they did not consider reperfusion treatment including IA thrombectomy related to prognosis.^{6,13,14,16-20} In addition, they were derived in high qualified health care system treating stroke of high-income developed countries, and they were developed based on the data from non-Asian patients.³⁻²⁰ However, mortality and disability after stroke were different between high-income countries and low- and middle-income countries. The prognostic score needs to be simple and easily applicable, using several variables related to outcome, validated externally, and needs to have good performance in the clinical setting.

Linkage with public healthcare data is useful in stroke research because stroke patients may enter into different sectors of the healthcare system before, during, and after stroke. Linked large-scale datasets enable researchers and healthcare practitioners to obtain a comprehensive view of stroke care and to improve national stroke care systems.^{21,22} In addition, large-scale linked administrative datasets are demonstrating increasing importance in epidemiological and clinical stroke research, since their use can improve research quality and transparency.^{23,24}

Therefore, we aimed to develop a good functional outcome prediction scoring system reflecting the current stroke managements for applying in the Korean patients including Asian stroke patients based on information available during hospitalization using a large dataset on stroke in Korea by linking the Clinical Research Center for Stroke (CRCS) registry with the Health Insurance Review and Assessment Service (HIRA) administrative claim database.

Methods

The CRCS data and the HIRA data

The CRCS registry was started in 2006 to collect data on acute stroke or transient ischemic attack (TIA) (within 7 days after onset) in Korea. The CRCS is supported by the Korea Healthcare Technology R&D Project of the Ministry of Health and Welfare in the Republic of Korea. Using a web-based database, CRCS collects clinical information on all acute stroke patients hospitalized at the neurology departments of a total of 65 participating hospitals. According to predefined protocols, demographic features, risk factors, stroke characteristics, treatment information, National Institute of Health Stroke Scale (NIHSS) scores, and laboratory information were collected at the time of entry into the registry database by stroke physicians or trained nurses. Data quality is monitored and audited regularly.²⁵⁻²⁷ The HIRA collected and managed claims data related to the National Health Insurance (NHI) program in the process of reimbursing healthcare providers in Korea. Accordingly, the HIRA database contains all information on the diagnoses, treatments, and prescribed medications for approximately 50 million Koreans. The information on prescribed drugs includes brand name, generic name, prescription date, duration of administration, and route of administration. In addition, all diagnoses are coded according to the International Classification of Disease, Tenth Revision (ICD-10).²⁸⁻³⁰

Data cleaning and preparation for linking datasets

We initially screened 108,430 stroke registry cases recorded from 65 participating hospitals in 2006 to 2014. These cases were screened based on the CRCS identifier. We excluded case records from 31 hospitals from which the patients were inconsistently registered ($n = 8,709$), patients who visited a hospital more than 7 days after stroke symptom onset ($n = 3,113$), and patients with lack of variables (insurance claim data) for linkage ($n = 31,297$). A total of 65,311 patients

from the hospitals were enrolled in the dataset that was used for linkage (Figure 1).

Data linkage methods

We linked CRCS and HIRA data (2007–2017) via a type of statistical matching that used common variables that were shared and stored in the enrolled hospitals and the HIRA. First, we used the claim data to identify common variables for linking the CRCS and the HIRA data. The selected common variables needed to be accurate, and there were no missing data for the linking process.³¹ From claim data, we chose four variables for matching: provider identification, receiving year, receiving number, and statement serial number. Additionally, we selected gender and date of birth as common variables for linking the two databases. Together, these six variables were used as the matching variables for the data linkage. The matching process was performed in the server for HIRA using Sybase IQ software (Sybase Inc., Dublin, CA, USA). After the linking was completed, all linked data were de-identified before analysis of the dataset.

Analysis of linkage accuracy and statistical analysis using linked data

First, we assess the matching rate (1:1 matching) of CRCS data that had been linked to HIRA data. Second, we evaluated linkage quality and errors during the linkage process. To assess the quality of the linked data, we compared the hospital visiting date in the CRCS data to the commencement date for health insurance care in the HIRA. If the difference between the two dates was 7 or fewer days, we accepted the case as a true match. In addition, we used absolute standardized differences (ASD) to compare the baseline characteristics of true matches and false matches in linked data. ASD analysis was used because it is expected to be more informative than P values for comparing large linked datasets, and may help to identify variables affected by potential bias due to linkage errors.^{23,32,33} Finally, we created a linked CRCS-HIRA database based on the truly

matched cases. The purpose of this database was to allow analyses of outcomes after index stroke.

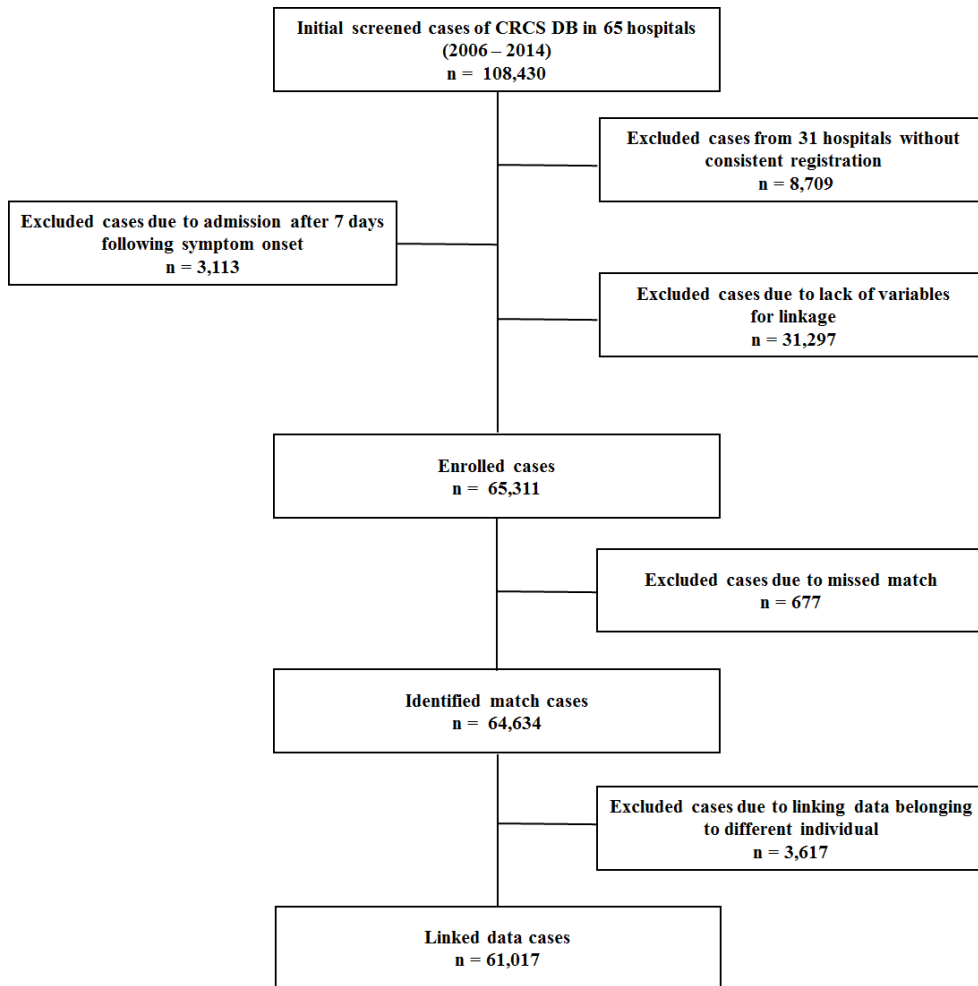


Figure 1. Flow diagram of establishing linkage dataset

CRCS = Clinical Research Center for Stroke

Derivation and validation groups

Over five consensus meetings with 16 clinical experts, we performed systematic reviews of the literatures and selected variables related to predicting functional outcomes. We also established inclusion criteria for the present study and set the derivation and validation groups for the development of a prediction score. The study was approved by the Institutional Review Boards (IRB) of Seoul National

University Hospital, 34 participating hospitals and HIRA (IRB No. H-1608-078-785). Informed consent was waived by the IRB.

The prognostic score was developed in the CRCS registry that collected clinical data from patients with acute strokes or transient ischemic attacks within 7 days of onset.^{25,27,34} We initially screened linking CRCS registry data with HIRA data ($n = 61,017$) between January 2007 and December 2014.³⁴ We included ischemic stroke patients ($n = 52,213$) in the linked data. We excluded patients who visited the hospital before July 2007 ($n = 2,511$), patients without 3-month modified Rankin scale (mRS) scores ($n = 24,734$), and patients with missing data regarding stroke mechanism ($n = 61$). Finally, we included 24,907 patients for development and external validation of the prognostic score. The derived prognostic score was also validated externally in the two independent (geographic and temporal differences) groups.³⁵⁻³⁷ Among the total included patients, we selected 22,005 patients for derivation of the prognostic score. The 2,902 patients from three centers with median value in baseline characteristics were identified as the external geographic validation group. The 531 patients with acute ischemic stroke were prospectively enrolled to reflect current status of stroke treatment such as increased IA thrombectomy from 5 centers between January 2018 and March 2018 as the externally temporal validation group for comparing prediction power during a different period. We divided patients into two groups with favorable outcome (mRS score ≤ 2) and unfavorable outcome (mRS score ≥ 3).

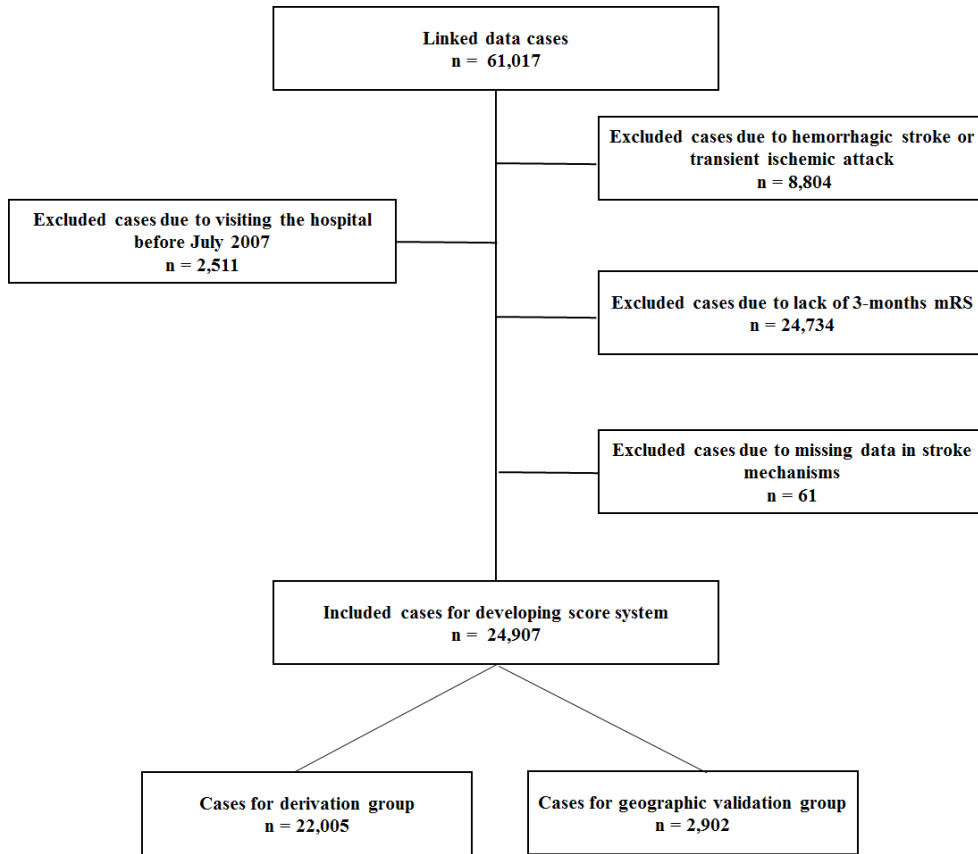


Figure 2. Flow diagram of included cases for developing prediction score system

Variables for the prediction scoring system

Several variables were obtained from the linked claim databases in the HIRA, including demographic factors, risk factors, stroke mechanisms, medications, reperfusion treatment, and stroke severity were collected from the CRCS registry, as well as comorbidity information, including dialysis and cancer. Data on risk factors (hypertension, diabetes mellitus, and hyperlipidemia) were updated using linked data. The history of risk factors (hypertension, diabetes mellitus, and hyperlipidemia) were defined as the use of anti-hypertensives, antidiabetics, and anti-hyperlipidemic medications with associated ICD-10 codes within 6 months before ischemic stroke in the linked claims data. The history of risk factors (atrial fibrillation, coronary

artery disease, and congestive heart failure) was defined using ICD-10 codes in the claims data within 6 months prior to ischemic stroke. Finally, we updated pre-stroke medications, including statins, antiplatelet agents, and anticoagulants, using information of prescribed medication taken from claims data in the linked data within 6 months of admission. Good 3-month outcome was defined as mRS 0-2, and poor outcome was defined as mRS 3-6.

We initially selected 22 possible predictive variables from systematic literature reviews and consensus meetings. Among them, 16 predictors associated with good outcome were selected using logistic regression analysis with stepwise backward elimination for developing the full version prediction score system. We categorized continuous variables for deriving the scoring system as follows: the age was dichotomized into age < 80, age ≥ 80 ^{38,39}; body mass index (BMI, kg/m²) was categorized into three groups, < 18.5, 18.5–24.9, and ≥ 25 ⁴⁰⁻⁴²; stroke severity was evaluated using the National Institutes of Health Stroke Scale (NIHSS), trichotomized into NIHSS 0–7, NIHSS 8–13, and NIHSS ≥ 14 ^{43,44}; and fasting plasma glucose (FPG) was dichotomized into FPG < 110 and FPG ≥ 110 .^{45,46} Stroke subtypes were classified as small vessel occlusion (SVO), cardioembolism (CE), and others (large artery atherosclerosis, other determined, and undetermined) based on the Trial of Org 10172 in Acute Stroke Treatment (TOAST) classification.⁴⁴ Among 16 predictors of full version scoring system, we selected 6 variables with high predictive power related to good outcome for developing an applicable scoring system. Predictors of simple version scoring system were decided through stroke experts' meetings.

Comparing developed score with other prediction score

To evaluate the use of available ischemic stroke outcome prediction scores, we compared the developed score with the THRIVE score^{5,6,48} for which sufficient data were available in our linked data set. We compared the performance of prediction at

3 months with THRIVE score, applying THRIVE scores to our linked registry data.^{6,48}

Statistical analysis

Baseline characteristics were presented as numbers (%) and continuous variables with normal distributions were presented as means \pm SD, while other variables that were not normally distributed were presented as medians (IQR). We used absolute standardized differences (ASD) to compare the baseline characteristics. ASD analysis was used because it is expected to be more informative than P-values for comparing large linked datasets.^{49,50} There was no multicollinearity among candidate variables for prediction. Logistic regression analysis with a fast-backward elimination method using the Akaike information criterion (AIC) was performed in the derivation cohort to identify predictors of favorable outcome (3-month mRS 0-2) after ischemic stroke. The prediction model was developed using logistic regression analysis with maximum likelihood estimation (MLE). Scoring system performance was assessed by the area under the receiver operating characteristic (AUC) curve (equivalent to the c statistic). Calibration was assessed with the Hosmer-Lemeshow test and calibration plots. Calibration plots were generated of predicted probability of good outcome versus actual probability of good outcome to assess model performance. The calibration slope is ideally equal to 1 and describes the effect of the predictors in the validation group versus those in the derivation group. The prediction score was developed from multiple logistic regression models using the regression coefficient-based scoring method. The total score was calculated by adding scores of each predictor. We performed external validations. External validation of the regression model between parameters of the S-SMART score and 3-month outcome was based on at least 1,000 bootstrap replicates. All statistical analyses were conducted by a professional medical statistician, J. S. Lee) using SAS 9.4 (SAS Institute, Inc. Cary, NC).

Results

Accuracy of data linkage between CRCS and HIRA data

A total of 65,311 cases were processed using the linkage algorithm, of which 677 cases were unmatched or one-to-many (1:M) matched. In total, 64,634 cases were one-to-one (1:1) matched in the HIRA dataset; the overall matching rate was 99.0%. As described in the Methods, we classified matches as true or false based on the difference between the hospital visiting date in the CRCS data and the commencement date for health insurance care in the HIRA data. Among the matched records, 61,017 cases (94.4%) were belonging to the same individual and 3,617 cases (5.6%) were belonging to the different individual, giving an accuracy rate in the total matched dataset of 94.4% (Figure 1). The baseline characteristics of true matches and false matches are summarized in Table 1. When we used ASD values to compare the baseline characteristics of true matched cases and false matched cases, no substantial difference was observed for any variable (Table 1).

The characteristics of the true matches were analyzed in detail. The mean age was 66.4 years and 58.4% of the patients were men. Recanalization treatments were received in 13.9% of the cases (intravenous [IV] thrombolysis 67.5%, endovascular treatment 16.0%, and combined IV thrombolysis and endovascular treatment 16.5%) and the median NIHSS score was 3 (IQR 1-7). Of the cases, 91.1% (n = 52,213) were ischemic stroke, 7.0% (n = 3,988) were TIA, and 1.9% (n = 1,113) were hemorrhagic stroke. Among the cases of ischemic stroke, 34.9% were accounted for by large artery atherosclerosis, 24.2% by SVO, 18.6% by CE, 2.6% by other determined factors, and 19.7% by undetermined factors.

Table 1. Baseline characteristics of matched cases according to linkage status

Variables	True matches (n = 61,017)	False matches (n = 3,617)	ASD
Age, mean (SD), y	66.4 ± 12.7	66.7 ± 11.8	0.023
Sex, male, n (%)	35,631 (58.4)	1,990 (55.0)	0.068
HT, n (%)	42,934 (70.4)	2,497 (69.0)	0.029
DM, n (%)	20,411 (33.5)	1,221 (33.8)	0.006
HL, n (%)	17,805 (29.2)	846 (23.4)	0.132
Previous Stroke/TIA, n (%)	10,662 (17.5)	759 (21.0)	0.089
Coronary heart disease, n (%)	4,544 (7.4)	177 (4.9)	0.106
A. fib, n (%)	10,592 (17.4)	526 (14.5)	0.077
Smoking, n (%)	23,720 (38.9)	1,196 (33.1)	0.121
Initial NIHSS, median (IQR)	3 (1 - 7)	3 (1 - 7)	0.021
Types of stroke, n (%)			
Ischemic stroke	52,213 (91.1)	3,020 (91.1)	0.001
Hemorrhagic stroke	1,113 (1.9)	100 (3.0)	0.069
TIA	3,988 (7.0)	194 (5.9)	0.045
Stroke mechanisms, n (%)			
LAA	18,236 (34.9)	1,018 (33.7)	0.026
SVO	12,617 (24.2)	784 (26.0)	0.041
CE	9,736 (18.6)	535 (17.7)	0.024
Other determined	1,337 (2.6)	83 (2.7)	0.012
Undetermined	10,287 (19.7)	600 (19.9)	0.004
Recanalization treatment, n (%)	8,457 (13.9)	347 (9.6)	
IV thrombolysis	5,706 (67.5)	221 (63.7)	0.122

IA thrombectomy	1,353 (16.0)	76 (21.9)	0.008
Combined IV thrombolysis and IA thrombectomy	1,398 (16.5)	50 (14.4)	0.068

ASD: Absolute standardized difference, SD: standard deviation, HT: hypertension, DM: diabetes mellitus, HL: hyperlipidemia, A.fib: atrial fibrillation, NIHSS: national institute of health stroke scale. IQR: interquartile range, TIA: transient ischemic attack, LAA: large artery atherosclerosis, SVO: small vessel occlusion, CE: cardioembolism, IV: intravenous, IA: intraarterial

Baseline characteristics of the derivation and validation groups

The demographic and baseline characteristics of patients included in the derivation group (n = 22,005), geographic validation group (n = 2,902) and temporal validation group are presented in Table 2 and Table 3. The patients in the derivation group were more likely to have atrial fibrillation and higher FPG levels. CE in stroke subtype was significantly higher and proportion of patients with no prestroke disability was lower in the derivation group. Other variables were similar between groups (Table 2). When comparing the baseline characteristics of the derivation group to those of the temporal validation group, the patients in the temporal validation group were older and more likely to have cancer; the proportions of hypertension, diabetes mellitus, hyperlipidemia, previous stroke or TIA, and use of antiplatelet agents or anticoagulants before stroke were significantly higher in the derivation group (Table 3). In the derivation group, 67.0% (n = 14,748) patients had good outcome and in the geographic validation group, 68.4% (n = 1,986) had good outcome at 3 months. In the temporal validation, good outcome rate at 3 month was 67.9% (n = 361), similar to the derivation and the geographic validation groups.

Table 2. Baseline characteristics of derivation and the geographic validation groups

	Derivation groups (n=22,005)	Geographic validation groups (n=2,902)	ASD
Age, mean (SD), y	67.1 ± 12.8	66.8 ± 13.2	0.021
Age, n (%)			0.029
< 80, y	18,553 (84.3)	2,416 (83.3)	
≥ 80, y	3,452 (15.7)	486 (16.7)	
Sex, male, n (%)	13,060 (59.4)	1,663 (57.3)	0.042
BMI, mean (SD)	23.7 ± 3.3	23.8 ± 3.4	0.022
BMI, n (%)			
<18.5	1,017 (4.6)	160 (5.5)	0.041
18.5-24.9	14,293 (65.0)	1,823 (62.8)	0.045
≥25	6,695 (30.4)	919 (31.7)	0.027
Initial NIHSS, median (IQR)	3.0 (1.0 - 7.0)	3.0 (1.0 - 7.0)	0.016
Initial NIHSS, n (%)			
0-7	16,721 (76.0)	2,263 (78.0)	0.047
8-13	2,699 (12.3)	315 (10.9)	0.044
≥14	2,585 (11.7)	324 (11.2)	0.018
Onset to ER visit time, median (IQR), h	8.6 (2.3 - 31.2)	9.7 (2.3 - 35.1)	0.043
Previous mRS = 0, n (%)	16,546 (75.2)	2,371 (81.7)	0.159
Stroke mechanisms, n (%)			
SVO	4,376 (19.9)	725 (25.1)	0.123
CE	4,654 (21.2)	461 (16.0)	0.136
Others	12,914 (58.8)	1,704 (59.0)	0.002
Previous stroke/TIA, n (%)	6,408 (29.1)	726 (25.0)	0.093
HT, n (%)	16,794 (76.3)	2,156 (74.3)	0.047

DM, n (%)	8,011 (36.4)	1,095 (37.7)	0.028
HL, n (%)	9,508 (43.2)	1,168 (40.2)	0.060
A.fib, n (%)	4,709 (21.4)	504 (17.4)	0.102
CHF, n (%)	1,826 (8.3)	254 (8.8)	0.016
Smoking, n (%)	8,823 (40.1)	1,305 (45.0)	0.099
Dialysis, n (%)	213 (1.0)	25 (0.9)	0.011
Cancer, n (%)	1,046 (4.8)	124 (4.3)	0.023
Pre-stroke antiplatelet agents/anticoagulants, n (%)	9,835 (44.7)	1,186 (40.9)	0.077
Pre-stroke statin, n (%)	5,714 (26.0)	637 (22.0)	0.094
FPG, mean (SD), mg/dL	119.6 ± 48.3	111.8 ± 40.2	0.177
FPG, n (%)			0.200
FPG < 110	12,695 (57.7)	1,954 (67.3)	
FPG ≥ 110	9,310 (42.3)	948 (32.7)	
Recanalization treatment, n (%)			
IV thrombolysis	1,973 (9.0)	233 (8.0)	0.034
IA thrombectomy	549 (2.5)	40 (1.4)	0.081
Combined IV thrombolysis and IA thrombectomy	794 (3.6)	100 (3.4)	0.009

ASD: Absolute standardized difference, SD: standard deviation, BMI: body mass index, ER: emergency room, NIHSS: national institute of health stroke scale, HT: hypertension, DM: diabetes mellitus, HL: hyperlipidemia, A.fib: atrial fibrillation, CHF: congestive heart failure, IQR: interquartile range, mRS: modified Rankin scale, TIA: transient ischemic attack, SVO: small vessel occlusion, CE: cardioembolism, FPG: fasting plasma glucose, IV: intravenous, IA: intraarterial

Table 3. Baseline characteristics of derivation and temporal validation group

	Derivation groups (n=22,005)	Temporal validation groups (n=531)	ASD
Age, mean (SD), y	67.1 ± 12.8	69.3 ± 13.3	0.175
Age, n (%)			0.181
< 80, y	18,553 (84.3)	410 (77.2)	
≥ 80, y	3,452 (15.7)	121 (22.8)	
Sex (male), n (%)	13,060 (59.4)	293 (55.2)	0.084
BMI, mean (SD)	23.7 ± 3.3	23.6 ± 3.7	0.038
BMI, n (%)			
< 18.5	1,017 (4.6)	37 (7.0)	0.101
18.5-24.9	14,293 (65.0)	336 (63.3)	0.035
≥ 25	6,695 (30.4)	158 (29.8)	0.015
Initial NIHSS, median (IQR)	3.0 (1.0 - 7.0)	4.0 (1.0 - 8.0)	0.051
Initial NIHSS, n (%)			
0-7	16,721 (76.0)	398 (75.0)	0.024
8-13	2,699 (12.3)	68 (12.8)	0.016
≥14	2,585 (11.7)	65 (12.2)	0.015
Onset to ER visit time, median (IQR), h	8.6 (2.3 - 31.2)	9.5 (2.0 - 34.8)	0.068
Previous mRS = 0, n (%)	16,546 (75.2)	392 (73.8)	0.031
Stroke mechanisms, n (%)			
SVO	4,376 (19.9)	110 (20.7)	0.019
CE	4,654 (21.2)	120 (22.6)	0.034
Others	12,914 (58.8)	301 (56.7)	0.044
Previous stroke/TIA, n (%)	6,408 (29.1)	128 (24.1)	0.114
HT, n (%)	16,794 (76.3)	331 (62.3)	0.307
DM, n (%)	8,011 (36.4)	160 (30.1)	0.133
HL, n (%)	9,508 (43.2)	167 (31.5)	0.245

A.fib, n (%)	4,709 (21.4)	122 (23.0)	0.038
CHF, n (%)	1,826 (8.3)	45 (8.5)	0.006
Dialysis, n (%)	213 (1.0)	11 (2.1)	0.090
Cancer, n (%)	1,046 (4.8)	55 (10.4)	0.213
Pre-stroke antiPLT/anticoagulant, n (%)	9,835 (44.7)	209 (39.4)	0.108
FPG, mean (SD), mg/dL	119.6 ± 48.3	120.4 ± 45.9	0.016
FPG, n (%)			0.039
FPG < 110	12,695 (57.7)	296 (55.7)	
FPG ≥ 110	9,310 (42.3)	235 (44.3)	
Recanalization treatment, n (%)			
IV thrombolysis	1,973 (9.0)	39 (7.3)	0.059
IA thrombectomy	549 (2.5)	24 (4.5)	0.110
Combined IV thrombolysis and IA thrombectomy	794 (3.6)	16 (3.0)	0.033

SD: standard deviation, BMI: body mass index, ER: emergency room, NIHSS: national institute of health stroke scale, HT: hypertension, DM: diabetes mellitus, HL: hyperlipidemia, A.fib: atrial fibrillation, CHF: congestive heart failure, IQR: interquartile range, mRS: modified Rankin scale, TIA: transient ischemic attack, SVO: small vessel occlusion, CE: cardioembolism, FPG: fasting plasma glucose, IV: intravenous, IA : intraarterial

Prediction scoring system after ischemic stroke

In the full version scoring system, 16 variables were identified as independent predictors of good outcome on multiple logistic regression analysis: age, gender, BMI, initial NIHSS, pre-stroke mRS, stroke mechanisms, previous stroke/TIA, hypertension, diabetes mellitus, hyperlipidemia, congestive heart failure, dialysis, cancer, pre-stroke antiplatelet agent or anticoagulant use, level of FPG, and recanalization treatment (only intravenous thrombolysis, IA thrombectomy,

combined IV thrombolysis and IA thrombectomy). The sum of the weighted scores was used to estimate the overall score for predicting good outcome. The total score of full-version scoring system was 252, and the optimal cut-off value for good outcome was 140 (Table 4). The performance of the full version prediction score system based on the AUC was 0.823 (0.817–0.829) in the derivation group (Figure 2). We developed the S-SMART score as simple version score for usefulness in clinical field using six variables: stroke **S**everity (NIHSS), **S**ex, stroke **M**echanism, **A**ge, pre-stroke **mRS**, and **T**hrombolysis/thrombectomy treatment. The total S-SMART score was 34 points, and optimal cut-off value related to good outcome was 17 points (Table 5). In the derivation group, the AUC of the S-SMART score was 0.805 (0.798–0.811), and the prediction power was comparable to the performance of full-version scoring system (Figure 3). Increasing scores predict an increasing chance of good outcome at 3 months after ischemic stroke (Figure 4).

Table 4. Full version scoring system for prediction of 3-month outcome in ischemic stroke patients

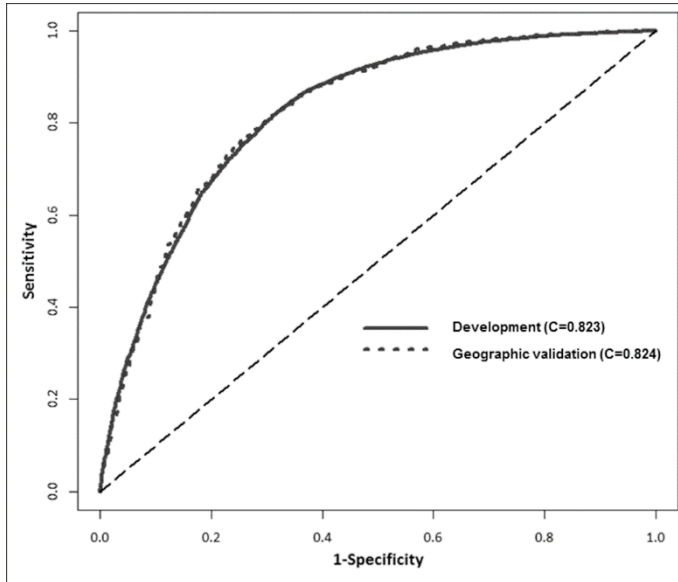
	Coef	S.E	Wald Z	Pr(> Z)	Score
Intercept	-5.915	0.223	-26.58	<0.001	
Age<80	1.048	0.047	22.32	<0.001	22
Male	0.471	0.036	13.20	<0.001	10
Initial NIHSS					
0-7	2.891	0.067	42.97	<0.001	62
8-13	0.929	0.071	13.08	<0.001	20
≥14	Ref				0
BMI					
<18.5	Ref	0.082	4.40	<0.001	0
18.5-24.9	0.361	0.087	6.27	<0.001	8
≥25	0.543				12
Pre-stroke mRS=0	0.691	0.040	17.24	<0.001	15

Stroke mechanisms					
SVO	0.476	0.049	9.69	<0.001	10
CE	0.290	0.047	6.12	<0.001	6
Others	Ref				0
Previous stroke=No	0.440	0.043	10.12	<0.001	9
HL=Yes	0.262	0.037	7.02	<0.001	6
Cancer=No	0.545	0.077	7.11	<0.001	12
DM=No	0.260	0.039	6.64	<0.001	6
HT=No	0.349	0.045	7.72	<0.001	7
CHF=No	0.179	0.065	2.77	0.006	4
Dialysis =No	0.576	0.161	3.57	<0.001	12
Pre-stroke					
antiPLT/anticoagulation=	0.047	0.042	1.13	0.261	1
Yes					
FPG < 110	0.409	0.037	11.01	<0.001	9
Recanalization treatment					
No	Ref				0
IV thrombolysis	0.180	0.065	2.79	0.005	4
IA thrombectomy	0.188	0.114	1.66	0.098	4
Combined IV		0.093	6.66	<0.001	
thrombolysis and IA	0.621				13
thrombectomy					
Total score					252

NIHSS: national institute of health stroke scale, BMI: body mass index, mRS: modified Rankin scale, HT: hypertension, DM: diabetes mellitus, HL: hyperlipidemia, CHF: congestive heart failure, TIA: transient ischemic attack, SVO: small vessel occlusion, CE: cardioembolism, FPG: fasting plasma glucose, IV: intravenous, IA: intraarterial

Figure 3. Receiver operating characteristic (ROC) curves of full version prediction score for functional outcome in derivation and external validation groups

A. ROC curves in derivation and geographic validation groups



B. ROC curves in derivation and temporal validation groups

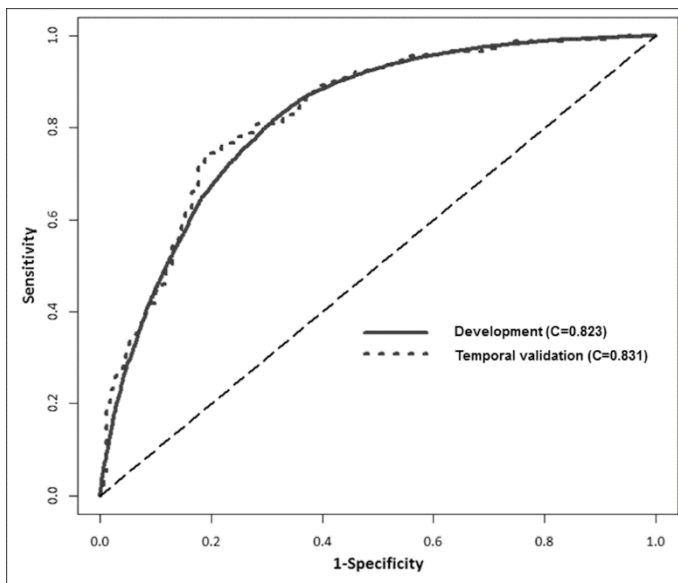


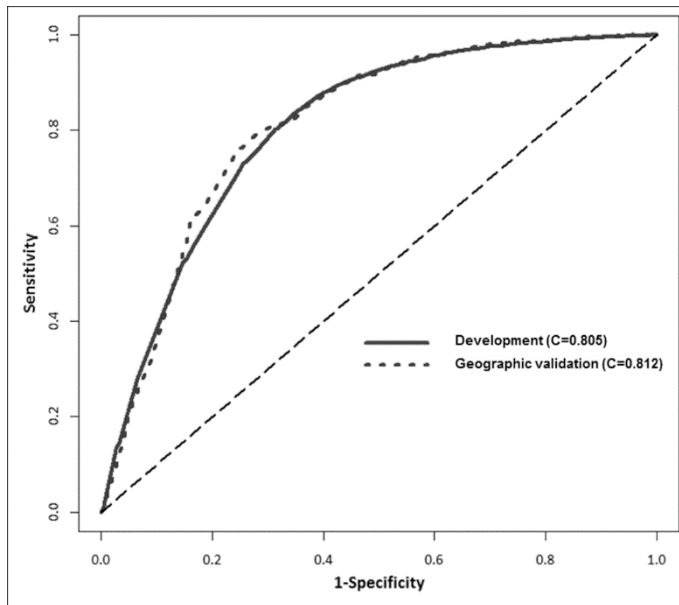
Table 5. S-SMART score for prediction of 3-month outcome in ischemic stroke patients

	Coef	S.E	Wald Z	Pr(> Z)	Score
Intercept	-3.534	0.0826	-42.78	<0.001	
Age<80	1.066	0.0454	23.51	<0.001	4
Male	0.461	0.0349	13.19	<0.001	2
Initial NIHSS					
0-7	2.946	0.0661	44.55	<0.001	12
8-13	0.950	0.0698	13.60	<0.001	4
≥14	Ref				0
Pre-stroke mRS=0	0.837	0.0381	21.94	<0.001	4
Stroke mechanisms					
SVO	0.501	0.0485	10.34	<0.001	2
CE	0.261	0.0452	5.77	<0.001	1
Others	Ref				0
Recanalization treatment					
No	Ref				0
IV thrombolysis	0.272	0.0634	4.28	<0.001	1
IA thrombectomy	0.238	0.1116	2.14	0.033	1
Combined IA thrombolysis and IA thrombectomy	0.713	0.0919	7.75	<0.001	3
Total score					34

NIHSS: national institute of health stroke scale, BMI: body mass index, mRS: modified Rankin scale, HT: hypertension, TIA: transient ischemic attack, SVO: small vessel occlusion, CE: cardioembolism, FPG: fasting plasma glucose, IV: intravenous, IA: intraarterial

Figure 4. Receiver operating characteristic (ROC) curves of the S-SMART score for functional outcome in derivation and external validation groups

A. ROC curves in derivation and geographic validation groups



B. ROC curves in derivation and temporal validation groups

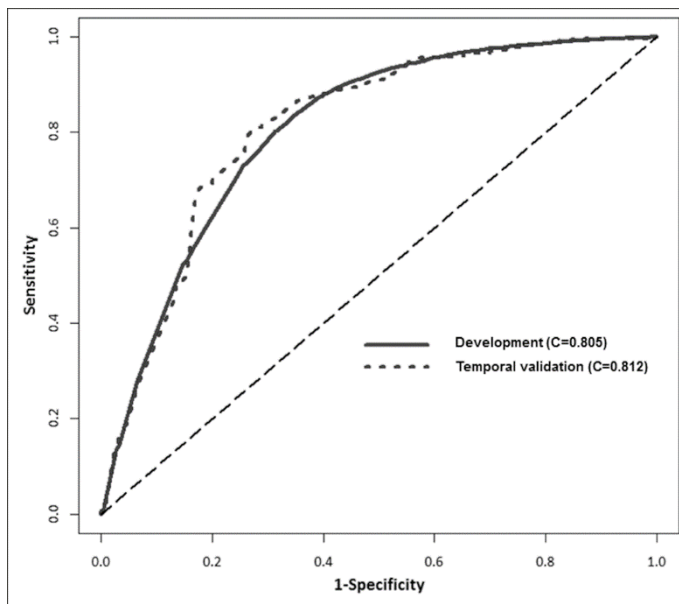
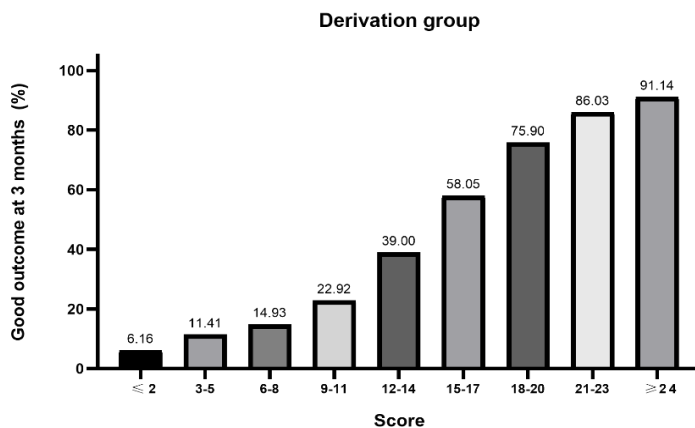


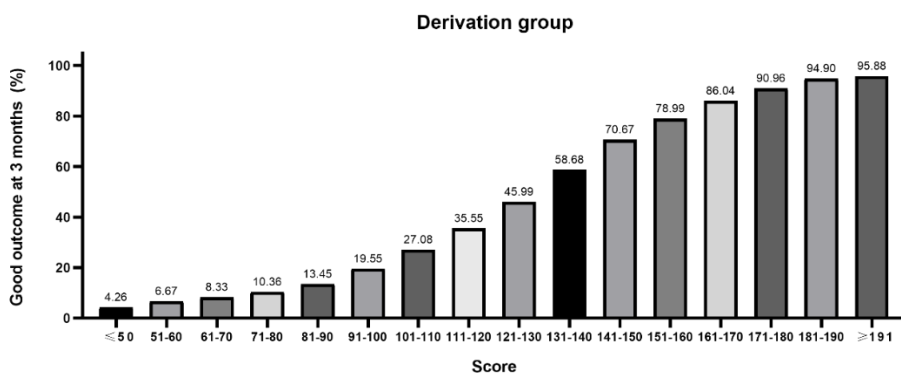
Figure 5. Association of prediction scores with good functional outcome

A. Association of S-SMART score with good functional outcome



A progressively higher percentage of patients with good functional outcome at 3 months is seen with each increased S-SMART score.

B. Association of full version score with good functional outcome



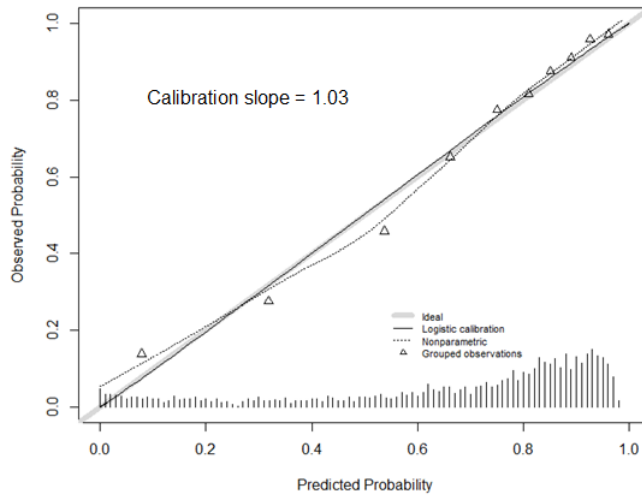
A progressively higher percentage of patients with good functional outcome at 3 months is seen with each increased prediction score.

Validation and model performance of the S-SMART score

Performance of the S-SMART score was validated in two external validation groups. The AUCs of the S-SMART scores were 0.812 (0.795–0.830) in the geographic validation group and 0.812 (0.771–0.854) in the temporal validation group for good outcome at 3 months, respectively (Figure 3). The calibration slopes were 1.03 and 0.86, respectively (Figure 5). When the full-version scoring system was applied to the two external validation groups, the AUCs were 0.824 (0.807–0.841) in the geographic validation group and 0.831 (0.793–0.869) in the temporal validation group (Figure 2). The calibration slopes of the full-version scoring system were 1.01 and 0.87 (Figure 6). Both scores showed good performance to predict outcome. After calibration tests, we updated the developed score system, and the calibration slope approached to 1.00. We compared the prediction power of the S-SMART score to that of the THRIVE score developed for prediction of functional outcome at 3 months. The AUC of full version scoring system showed higher value compared to the THRIVE score [0.855 (0.850–0.860) vs. 0.839 (0.833–0.844), $P < 0.001$]. The AUC of S-SMART score was significantly higher than that of the THRIVE score for prediction of outcome at 3 months [0.848 (0.843–0.854) vs. 0.839 (0.833–0.844), $P < 0.001$]. When comparing the calibration slopes of two scores, the slope of the S-SMART score in the external validation group (1.03) was closer to 1 than was THRIVE score in the derivation and geographic validation groups (1.16 and 1.18, respectively) (Figure 7). Therefore, the predictive power of the S-SMART score was superior or comparable to that of the THRIVE score.

Figure 6. Calibration plots of S-SMART score for predicting functional outcome in two external validation groups

A. Calibration test in the geographic validation group



B. Calibration test in the temporal validation group

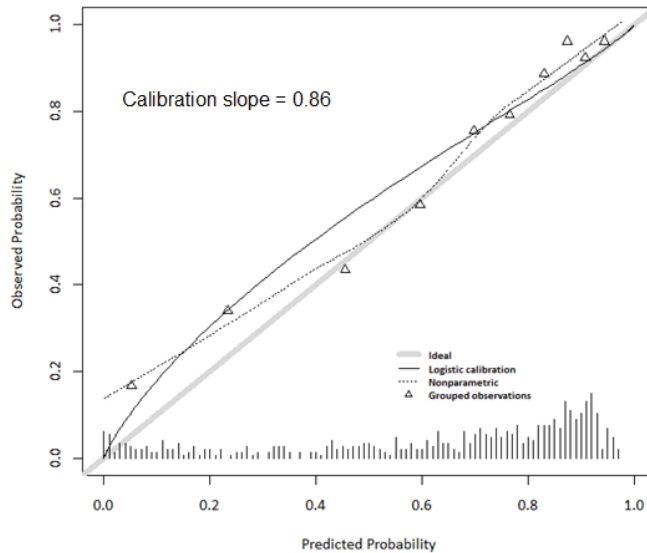
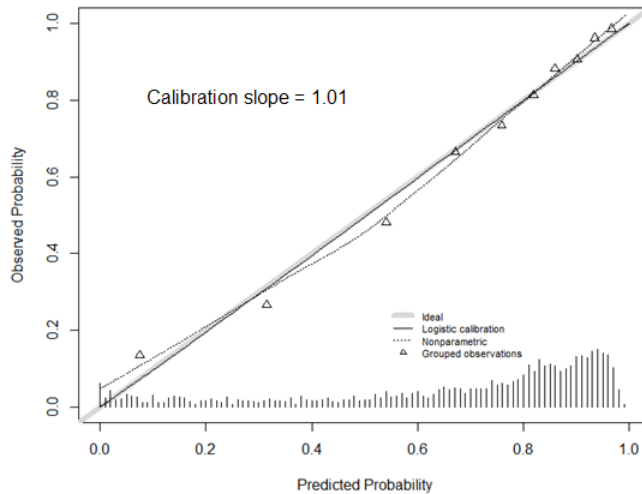


Figure 7. Calibration plots of full version score for predicting functional outcome in two external validation groups

A. Calibration test in the geographic validation group



B. Calibration test in the temporal validation group

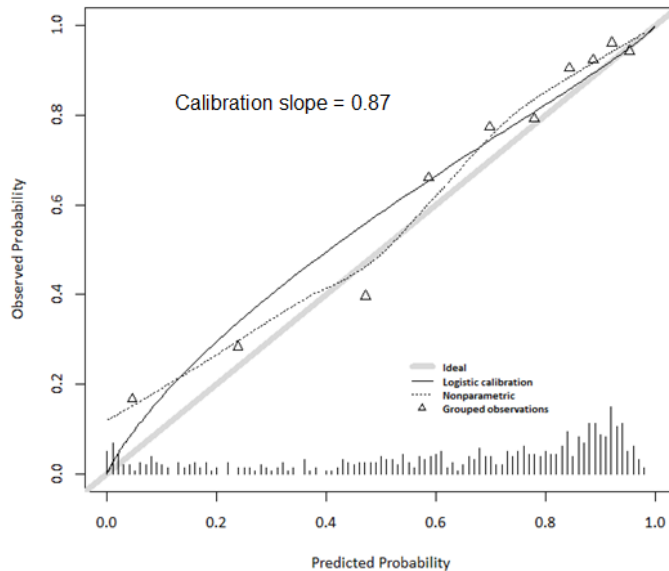
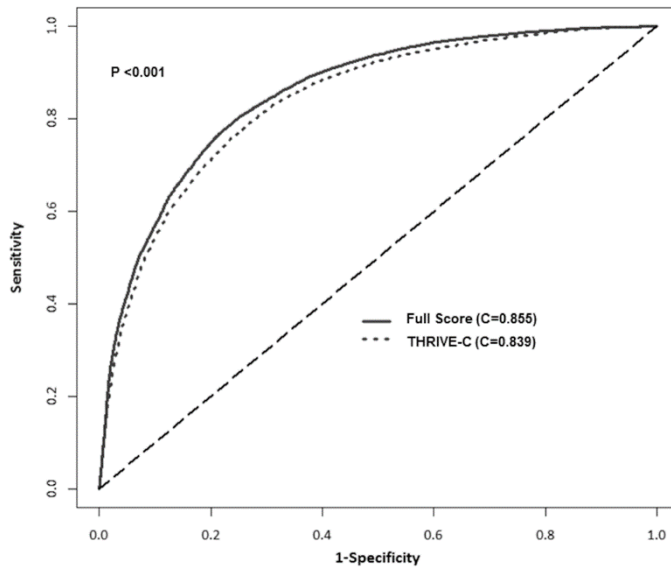
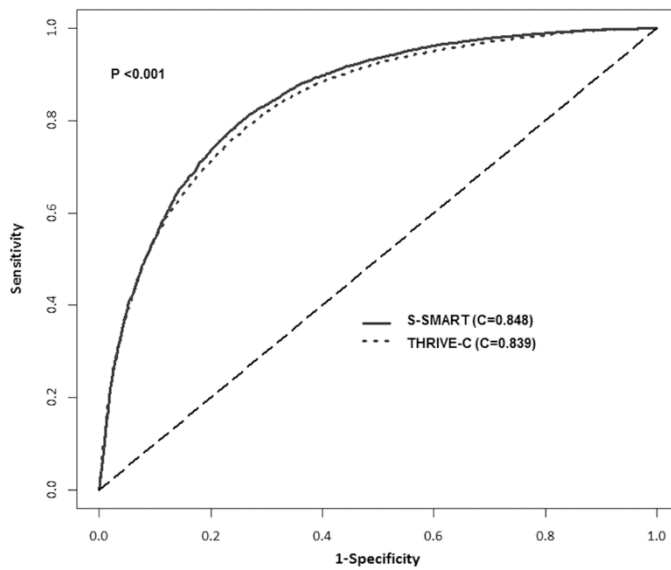


Figure 8. Comparing prediction power of S-SMART score to THRIVE score

A. ROC curves in full version score vs. THRIVE score

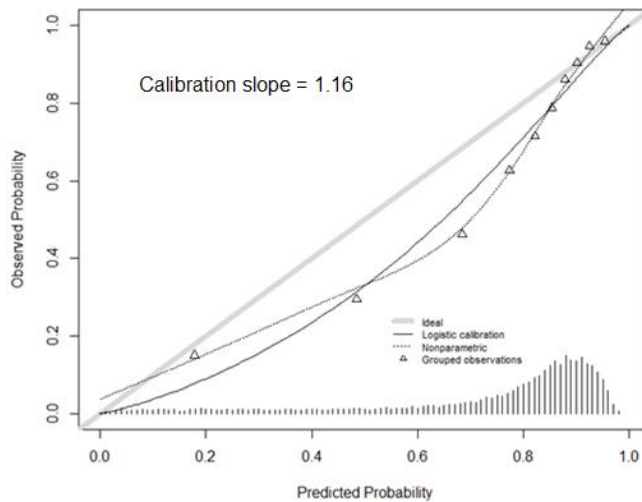


B. ROC curves in S-SMART score vs. THRIVE score

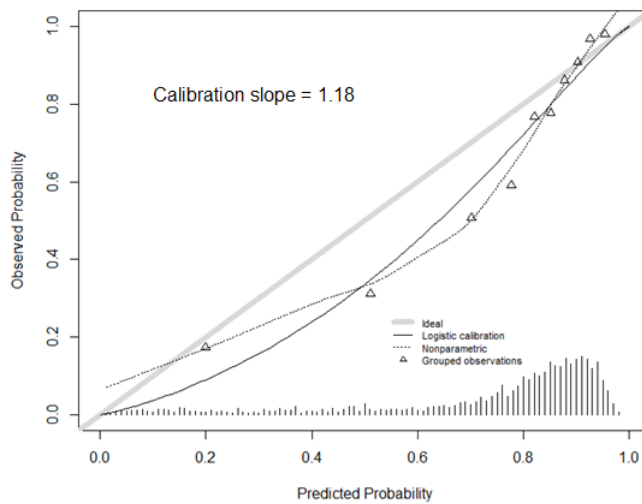


C. Calibration test of S-SMART score vs. THRIVE score

C-1. THRIVE score in derivation group



C-2. THRIVE score in geographic validation group



Discussion

The S-SMART score is a prognostic score that can assess good outcome at 3 months during the acute stage after ischemic stroke. The derived score consists of six variables (stroke severity, gender, stroke mechanism, pre-stroke mRS, thrombolysis/thrombectomy treatment) that can be assessed at the bedside during hospitalization. The score was developed based on the linked big data and well validated in two external groups with high performance. Moreover, S-SMART was found to be superior to other clinical prediction scores on direct comparison.

The S-SMART score includes objective tools (16 variables in the full model and six variables in the simple model) to stratify and estimates the risk of a good functional outcome after stroke. The score performed well on external validations, and can be applied in clinical practice as well as stroke research. The S-SMART score is a reliable score for predicting prognosis using linked big data for stroke. It can be used during the acute phase following ischemic stroke and it can be further developed through an additional external validation study. We will set a system of estimating probability of outcome according to scores based on the result of a future validation study. This prediction model may assist estimation of functional outcome after stroke and to determine care plans after stroke.

Several prognostic scores such as SPI-II score,⁷ modified SOAR score,⁹ PLAN score¹⁰, and iScore,¹³ after stroke have been developed for predicting mortality after stroke.^{7, 9,10,11,13,14} In addition, some scores including HAT score,⁸ SPAN-100 score,¹² and DRAGON score,¹⁵ were developed for predicting outcome in patients treated with IV thrombolysis. Moreover, the ASTRAL score predicted functional outcome at 3 months after stroke using neurological symptom, stroke severity, age, and onset to admission time (Table 6).¹⁷ Previous developed scores did not consider reperfusion treatment such as intra-arterial thrombectomy³⁻²⁰ or stroke mechanisms related outcome.⁴⁻¹⁵ Moreover, they were derived using stroke cohorts with high qualified health care system treating stroke of high-income developed

countries, especially in non-Asian patients.³⁻²⁰ Higher mortality and disability after stroke occurred frequently in low- and middle-income countries.⁵¹⁻⁵⁴

Table 6. Prognostic scores for predicting outcome in stroke patients

Scores	Derived Study populations	Outcome
HAT score ⁸	Patients with IS treated IV thrombolysis (NINDS cohort, USA)	Risk of hemorrhage
SPAN-100 score ¹²	Patients with IS treated IV thrombolysis (NINDS cohort, Canada and USA)	ICH and 3-month functional outcome
DRAGON score ¹⁵	Patients with IS treated IV thrombolysis (Helsinki University Central Hospital, Finland)	3-month functional outcome
THRIVE score ⁵	Patients in MERCI and multi MERCI trials (USA)	Outcome and mortality at 90 days
SOAR score ⁹	Acute stroke (UK stroke registry, UK)	Early mortality within 7 days and hospital length of stay
IScore ¹³	Patients with IS (RCSN, Canada)	30-day and 1-year mortality
PLAN score ¹⁰	Patients with TIA or IS (RCSN, Canada)	30-day and 1-year mortality and a modified Rankin score of 5 to 6 at discharge
SPI-II score ⁷	Patients with TIA or IS (WEST cohort, USA and UK)	Stroke or death in 2 years
ASTRAL score ¹⁷	Patients with IS (ASTRAL cohort, Europe)	3-month modified Rankin Scale score

IS: ischemic stroke, IV: intravenous, NINDS: National Institute of Neurological Disorders and Stroke, RCSN: Registry of the Canadian Stroke Network, TIA: transient ischemic attack

The S-SMART score has several advantages over more recently developed ischemic stroke prediction scores. One advantage is that it can be easily calculated during the acute stage at the bedside using a simple summation to predict good functional outcome. Furthermore, by contrast with previous scores, this score does not require brain imaging information or subjective information such as neurological examination results.^{15,17} This score also considers stroke mechanisms and reperfusion treatments as independent predictors of outcome. We selected the 3 months mRS as the outcome of score. The functional disability associated with daily life is more important than mortality for planning of long-term care after stroke for patients, families, and physicians. Moreover, it was derived from linked big data, based on large registry of consecutive ischemic stroke patients, and it performed well in external validation studies. We chose to validate the score in two independent cohorts at various centers and various temporal variations (time periods) reflecting variable clinical practice, in order to evaluate its validity and applicability in ischemic stroke patients with varying baseline characteristics. Therefore, our developed score used linked big data that could be applied to Asian patients. Finally, this score predicts functional outcomes across the full range of acute stroke management, including intravenous thrombolysis, recanalization treatment, and absence of hyperacute treatment.

Moreover, we have established a large dataset on stroke by linking the CRCS registry and administrative HIRA data. A 99.0% matching rate was achieved by using data from claims as matching variables, without relying on personal identifiers. Additionally, the accuracy of the linkage was high (94.4%). Therefore, linkage of the HIRA and clinical data, such as the CRCS data from hospitalizations, could serve as a powerful research resource to study stroke prognosis and healthcare service utilization, from acute to chronic stages of stroke.

There were some limitations in our study. First, although we controlled and updated several variables related to stroke outcome, we could not rule out the

possibility that additional baseline variables (unmeasured confounds) including stroke unit care related to outcome after stroke^{56, 57} may have some impact. Second, the score was developed in Korean stroke patients; therefore, evaluating the reliability in other population is needed. In addition, generalization of this score could be limited in low- and middle-income developing countries, because of lack of acute stroke management system, combined with inadequate rehabilitation services, and lack of preventive measures in stroke patients.⁵¹⁻⁵⁴ Third, participating centers in the CRCS registry were large centers, therefore there could be limitations related to generalization to community hospitals and small centers. Fourth, our linking method is only possible in studies with access to information from hospitals. We linked data using the common claim data in each hospital record and in the HIRA. Moreover, data linkage accuracy is dependent on the quality of matching variables. Therefore, it would be difficult to use our method for linking data from cohort studies that do not have access to claims. Despite these limitations, we built a linked, large data source on stroke in Korea, and developed the prognostic score for predicting functional outcome after stroke data with high predictive power using the linked big data.

The S-SMART score could be applied and validated in several countries including low- and middle-income developing countries and high qualified health care system treating stroke with stroke registry of Asia. In addition, it can be validated in the community hospitals and small centers for generalization in Korea. Moreover, we expect to perform several nationwide stroke studies, including epidemiological analyses, comprehensive assessments of the national stroke care system, and research directed at the goal of improving stroke care using this score and linked data.

The S-SMART score is an applicable prediction method during the acute stage after ischemic stroke that can be used when counseling patients and families. It performed well in two external validations and may be a useful tool for clinical

practice and stroke research. It not only provides an estimation of outcome, but it also can provide support regarding stroke management in the future for patients and their families. Moreover, it can be used in large clinical trials to select enrolled patients. Furthermore, S-SMART score could be improved for predicting long-term mortality and stroke recurrence based on the linking other big data from Statistics Korea and National Health Insurance Service data using identifier. Further external validation studies are needed in small centers and developed counties for generalization of S-SMART score.

Reference

1. E. Mayo, N., *et al.* Disablement following stroke. **21**, 258-268 (1999).
2. Nichols-Larsen, D.S., Clark, P., Zeringue, A., Greenspan, A. & Blanton, S.J.S. Factors influencing stroke survivors' quality of life during subacute recovery. **36**, 1480-1484 (2005).
3. Abdul-Rahim, A.H., *et al.* Derivation and Validation of a Novel Prognostic Scale (Modified-Stroke Subtype, Oxfordshire Community Stroke Project Classification, Age, and Prestroke Modified Rankin) to Predict Early Mortality in Acute Stroke. *Stroke* **47**, 74-79 (2016).
4. de Ridder, I.R., *et al.* Development and validation of the Dutch Stroke Score for predicting disability and functional outcome after ischemic stroke: A tool to support efficient discharge planning. **3**, 165-173 (2018).
5. Flint, A., Cullen, S., Faigeles, B. & Rao, V.J.A.J.o.N. Predicting long-term outcome after endovascular stroke treatment: the totaled health risks in vascular events score. **31**, 1192-1196 (2010).
6. Flint, A.C., *et al.* THRIVE score predicts ischemic stroke outcomes and thrombolytic hemorrhage risk in VISTA. **44**, 3365-3369 (2013).
7. Kernan, W.N., *et al.* The stroke prognosis instrument II (SPI-II) : A clinical prediction instrument for patients with transient ischemia and nondisabling ischemic stroke. *Stroke* **31**, 456-462 (2000).
8. Lou, M., *et al.* The HAT Score: a simple grading scale for predicting hemorrhage after thrombolysis. **71**, 1417-1423 (2008).
9. Myint, P.K., *et al.* The SOAR (Stroke subtype, Oxford Community Stroke Project classification, Age, prestroke modified Rankin) score strongly predicts early outcomes in acute stroke. **9**, 278-283 (2014).
10. O'donnell, M.J., *et al.* The PLAN score: a bedside prediction rule for death and severe disability following acute ischemic stroke. **172**, 1548-1556 (2012).

11. Park, T.H., *et al.* The iScore predicts clinical response to tissue plasminogen activator in Korean stroke patients. **23**, 367-373 (2014).
12. Saposnik, G., Guzik, A.K., Reeves, M., Ovbiagele, B. & Johnston, S.C.J.N. Stroke prognostication using age and NIH Stroke Scale: SPAN-100. **80**, 21-28 (2013).
13. Saposnik, G., *et al.* IScore: a risk score to predict death early after hospitalization for an acute ischemic stroke. *Circulation* **123**, 739-749 (2011).
14. Saposnik, G., *et al.* The iScore predicts poor functional outcomes early after hospitalization for an acute ischemic stroke. *Stroke* **42**, 3421-3428 (2011).
15. Strbian, D., *et al.* Predicting outcome of IV thrombolysis–treated ischemic stroke patients The DRAGON score. **78**, 427-432 (2012).
16. Quinn, T.J., *et al.* Validating and comparing stroke prognosis scales. 10.1212/WNL.0000000000004332 (2017).
17. Ntaios, G., *et al.* An integer-based score to predict functional outcome in acute ischemic stroke the ASTRAL score. WNL. 0b013e318259e318221 (2012).
18. Rha, J.-H. & Saver, J.L. The impact of recanalization on ischemic stroke outcome: a meta-analysis. *Stroke* **38**, 967-973 (2007).
19. Goyal, M., *et al.* Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *The Lancet* **387**, 1723-1731 (2016).
20. Emberson, J., *et al.* Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *The Lancet* **384**, 1929-1935 (2014).
21. Jutte, D.P., Roos, L.L. & Brownell, M.D.J.A.r.o.p.h. Administrative record linkage as a tool for public health research. **32**, 91-108 (2011).

22. Zingmond, D.S., Ye, Z., Ettner, S.L. & Liu, H.J.J.o.c.e. Linking hospital discharge and death records—accuracy and sources of bias. **57**, 21-29 (2004).
23. Kim, L., Kim, J.-A., Kim, S.J.E. & health. A guide for the utilization of health insurance review and assessment service national patient samples. **36**(2014).
24. Bradley, C.J., Penberthy, L., Devers, K.J. & Holden, D.J.J.H.s.r. Health services research and data linkages: issues, methods, and directions for the future. **45**, 1468-1488 (2010).
25. Kim, B.J., *et al.* Case characteristics, hyperacute treatment, and outcome information from the clinical research center for stroke-fifth division registry in South Korea. **17**, 38 (2015).
26. Park, T.H., *et al.* Gender differences in the age-stratified prevalence of risk factors in Korean ischemic stroke patients: a nationwide stroke registry-based cross-sectional study. **9**, 759-765 (2014).
27. Hong, K.-S., *et al.* Stroke statistics in Korea: part I. Epidemiology and risk factors: a report from the Korean stroke society and clinical research center for stroke. **15**, 2 (2013).
28. Kim, H.-A., *et al.* The epidemiology of total knee replacement in South Korea: national registry data. **47**, 88-91 (2008).
29. Shin, J.-Y., *et al.* Risk of ischemic stroke with the use of risperidone, quetiapine and olanzapine in elderly patients: a population-based, case-crossover study. **27**, 638-644 (2013).
30. Kim, J., Yoon, S., Kim, L.-Y. & Kim, D.-S.J.J.o.K.m.s. Towards actualizing the value potential of Korea Health Insurance Review and Assessment (HIRA) data as a resource for health research: strengths, limitations, applications, and strategies for optimal use of HIRA data. **32**, 718-728 (2017).
31. D’Orazio, M. Statistical Matching and Imputation of Survey Data with

StatMatch. (2017).

32. Austin, P.C.J.S.i.m. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. **28**, 3083-3107 (2009).
33. Ford, J.B., Roberts, C.L., Taylor, L.K.J.P. & Epidemiology, P. Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. **20**, 329-337 (2006).
34. Kim, T.J., *et al.* Building Linked Big Data for Stroke in Korea: Linkage of Stroke Registry and National Health Insurance Claims Data. **33**(2018).
35. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of internal medicine*. 1999;130(6):515-524.
36. Moons KG, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.
37. Steyerberg EW, Vergouwe YJEhj. Towards better clinical prediction models: seven steps for development and an ABCD for validation. 2014;35(29):1925-1931.
38. Di Carlo, A., *et al.* Stroke in the very old: clinical presentation and determinants of 3-month functional outcome: a European perspective. *Stroke* **30**, 2313-2319 (1999).
39. Saposnik, G., *et al.* Stroke outcome in those over 80: a multicenter cohort study across Canada. *Stroke* **39**, 2310-2317 (2008).
40. Aparicio, H.J., *et al.* Overweight, obesity, and survival after stroke in the Framingham Heart Study. *Journal of the American Heart Association* **6**, e004721 (2017).
41. Kawase, S., *et al.* Association between body mass index and outcome in J apanese ischemic stroke patients. *Geriatrics & gerontology international* **17**,

- 369-374 (2017).
42. Organization, W.H. Obesity: preventing and managing the global epidemic: report of a WHO consultation on obesity, Geneva, 3-5 June 1997. (Geneva: World Health Organization, 1998).
 43. Fonarow, G.C., *et al.* Relationship of National Institutes of Health Stroke Scale to 30-day mortality in Medicare beneficiaries with acute ischemic stroke. *Journal of the American Heart Association* **1**, e000034 (2012).
 44. Rubin, G., Firlik, A.D., Levy, E.I., Pindzola, R.R. & Yonas, H. Relationship between cerebral blood flow and clinical outcome in acute stroke. *Cerebrovascular Diseases* **10**, 298-306 (2000).
 45. Alberti, K.G.M.M. & Zimmet, P.f. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. *Diabetic medicine* **15**, 539-553 (1998).
 46. Committee, I.E. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes care* **32**, 1327-1334 (2009).
 47. Adams Jr, H.P., *et al.* Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke*; **24**, 35-41 (1993).
 48. Kamel, H., *et al.* The Total Health Risks in Vascular Events (THRIVE) score predicts ischemic stroke outcomes independent of thrombolytic therapy in the NINDS tPA trial. *Journal of Stroke and Cerebrovascular Diseases* ;**22**, 1111-1116 (2013).
 49. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, *et al.* A guide to evaluating linkage quality for the analysis of linked data. *International journal of epidemiology*. 2017;46(5):1699-1710.
 50. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples.

Statistics in medicine. 2009;28(25):3083-3107.

51. Sposato, L.A., *et al.* Quality of ischemic stroke care in emerging countries: the Argentinian National Stroke Registry (ReNACer). *Stroke* 2008;39: 3036-3041.
52. Johnson, W., Onuma, O., Owolabi, M. & Sachdev, S. Stroke: a global response is needed. *Bulletin of the World Health Organization* 2016; 94: 634
53. Bender, M., *et al.* High burden of stroke risk factors in developing country: The case study of Bosnia-Herzegovina. *Materia socio-medica*, 2017;29: 277.
54. Feigin, V.L. Stroke epidemiology in the developing world. *The Lancet* 2005; 365: 2160-2161.
55. Candelise, L., *et al.* Stroke-unit care for acute stroke patients: an observational follow-up study. *The Lancet*; 2007;369:299-305.
56. Collaboration, S.U.T. Organised inpatient (stroke unit) care for stroke. *Cochrane Database of Systematic Reviews* 2013.

국문초록

배경 및 목적: 뇌졸중환자의 급성기 데이터를 이용하여 뇌졸중 이후 기능적 예후를 예측하게 된다면 향후 환자의 치료, 효과적인 관리와 장기적인 계획을 세우는 데 도움이 될 수 있다. 빅데이터를 이용한 공공 보건 의료 데이터의 연계는 뇌졸중 전후의 환자 상태를 확인할 수 있어 뇌졸중 연구에 유용하다. 이에, 본 연구에서는 연계 빅데이터를 이용하여 허혈성 뇌졸중 후 기능적 예후를 예측하기 위한 점수체계 개발하고 검증하고자 하였다.

방법: 본 연구는 뇌졸중임상연구센터 (Clinical Research Center for Stroke, CRCS) 레지스트리에 2006년부터 2014년까지 등록된 환자들 중 빅데이터와 연계 가능한 급성 뇌경색 환자 65,311명의 자료를 이용하여 진행하였다. 6개의 공통 변수인 생년월일, 성별, 요양기호, 접수년도, 접수번호, 명세서일련번호를 이용하여 CRCS 레지스트리 자료와 건강보험심사평가원 (Health Insurance Review and Assessment Service, HIRA)의 청구자료를 연계하였다. 연계 데이터의 연계 정확도는 CRCS 레지스트리의 내원일자와 HIRA의 요양개시일 간의 차이를 이용하여 평가하였다. 연계된 자료 중 2007년 7월부터 2014년 12월까지 CRCS-HIRA 관련 데이터에서 급성 뇌경색 환자 22,005명을 예후 예측 모델 개발을 위한 집단으로 선정하고 연구를 수행하였다. 예후는 급성 뇌경색 발생 후 3개월째의 modified Rankin scale (mRS)을 사용하여 평가하였다. 좋은 예후군을 mRS 2점 이하인 군으로 정의하고 이와 관련된 예측 인자를 식별하고 로지스틱 회귀 계수를 사용하여 점수체계를 개발하였다. 본 연구에서 개발된 예후 예측 점수체계는 2개의 외적 타당도 평가 (지리적 타당도 평가군과 시간적

타당도 평가군)군에서 검증되었다. 예후 예측 점수체계의 예측력은 AUC(area under the plasma level-time curve)를 이용하여 평가되었다.

결과: 본 연구의 연계 빅데이터의 연계 정확도는 94.4% 였다. 이러한 연계 빅데이터를 기반으로 하여 뇌졸중 중증도, 성별, 뇌졸중 기전, 연령, 뇌졸중 전 mRS 및 혈전 용해 치료의 6개 변수가 뇌졸중 후 3개월 기능적 예후 예측 S-SMART 점수 체계 (총 34 점)의 예측 변수로 선정되었다. 예측 점수 체계의 AUC는 점수 체계 개발 군에서 0.805 (0.798-0.811)이었다. 이 모델의 AUC는 지리적 타당도 평가 군에서의 경우 0.812 (0.795-0.830) 이었고 시간적 타당도 평가 군에서의 경우 0.812 (0.795-0.830) 이었다.

결론: 본 연구에서는 CRCS 레지스트리와 HIRA 자료 연계를 통하여 뇌졸중 관련 빅데이터를 구축하였다. 또한, 연계자료를 이용하여 급성 뇌경색 이후 기능적 예후를 예측할 수 있는 S-SMART 점수 체계를 개발하였다. 본 예측 모델은 뇌졸중 후 예후 평가를 하여 뇌졸중 후 치료 계획을 결정할 수 있도록 도움을 줄 수 있을 것이다.

주요어: 빅데이터, 자료 연계, 뇌경색, 예후, 예측 점수

학 번: 2014-30679